Quantification of Profiles[1]

Ernest C. Davenport, Jr.
University of Minnesota
172 Education Sciences Building
56 East River Road
Minneapolis, MN 55455
LQR6576@umn.edu

Mark L. Davison
University of Minnesota
182 Education Sciences Building
56 East River Road
Minneapolis, MN 55455
MLD@umn.edu

1. Abstract

The use of profiles to determine disease states, apprehend criminals, identify successful candidates for school or work, etc. has been a staple in many clinical and applied fields (e.g. medicine, psychology, college admissions, human resources). Regardless of discipline, multiple measures are taken and the pattern of scores on these measures is compared to a prototypical pattern, where the prototypical pattern shows the pattern of scores indicative of a particular classification or diagnosis. While the measures provide quantitative data, the final decision is often based on expert opinion regarding the closeness of an individual's pattern of scores to that of the prototypical pattern. Thus, the final decision is oftentimes as much art as objective. This paper provides a procedure to quantify the match of score information for an individual with a prototypical pattern. In addition to matching a predetermined optimal pattern, we provide a two-step approach that first allows for identification of the optimal pattern, specified as a vector of contrast coefficients, and then provides quantification of the match to this optimal pattern. The procedure need not supplant clinical judgment, but can be used as one of the elements that enter into a clinician's decision.

2. Introduction

There has been much activity since the 1950's to quantify score profiles (Meehl, 1950, Cronbach and Gleser 1953). Most of the earlier methods yield profiles that need not have criterion-related validity. Indeed, cluster analysis, modal profile analysis, and profile analysis via multidimensional scaling rely solely on the subtest variables within each profile, and no external criterion is used to identify the core profile types. Thus, one cannot be sure the resulting profiles have criterion validity. Davison and Davenport (2002) suggest a procedure that quantifies the relationship of individual profiles to an external criterion. It makes use of Cronbach and Gleser's (1953) decomposition of a set of scores into three components: elevation, scatter, and shape. Specifically, an index that combines scatter and shape allows for a profile match statistic that quantifies a subject's score profile match to an empirical and/or theoretical prototypical profile, and this leads to an estimate of the criterion variance accounted for by the optimal pattern, or more precisely the variance accounted for by the match statistic. This match statistic will be demonstrated for mathematics course taking patterns.

3. Development of the Procedure

To fully motivate this discussion we start with the development of the procedure. Note that one can begin at step 11 (below) if the prototypical pattern is already specified; in that case one need only provide a match statistic to that existing pattern. Steps 1 – 10 justify our approach to identify an optimally predictive pattern and then show how we can quantify a match to that pattern. Specifically, we will show how the original regression equation can be parsed into *Level* and *Pattern* components based on a least squares fit to the criterion. The resulting pattern is optimal via that least squares fit. Once the optimal pattern is defined we then provide a statistic that will quantify each observed profile's match to that pattern and the amount of criterion variance accounted for by the match statistic.

---

[1] 2012 Federal Committee on Statistical Methodology Research Conference, Washington DC

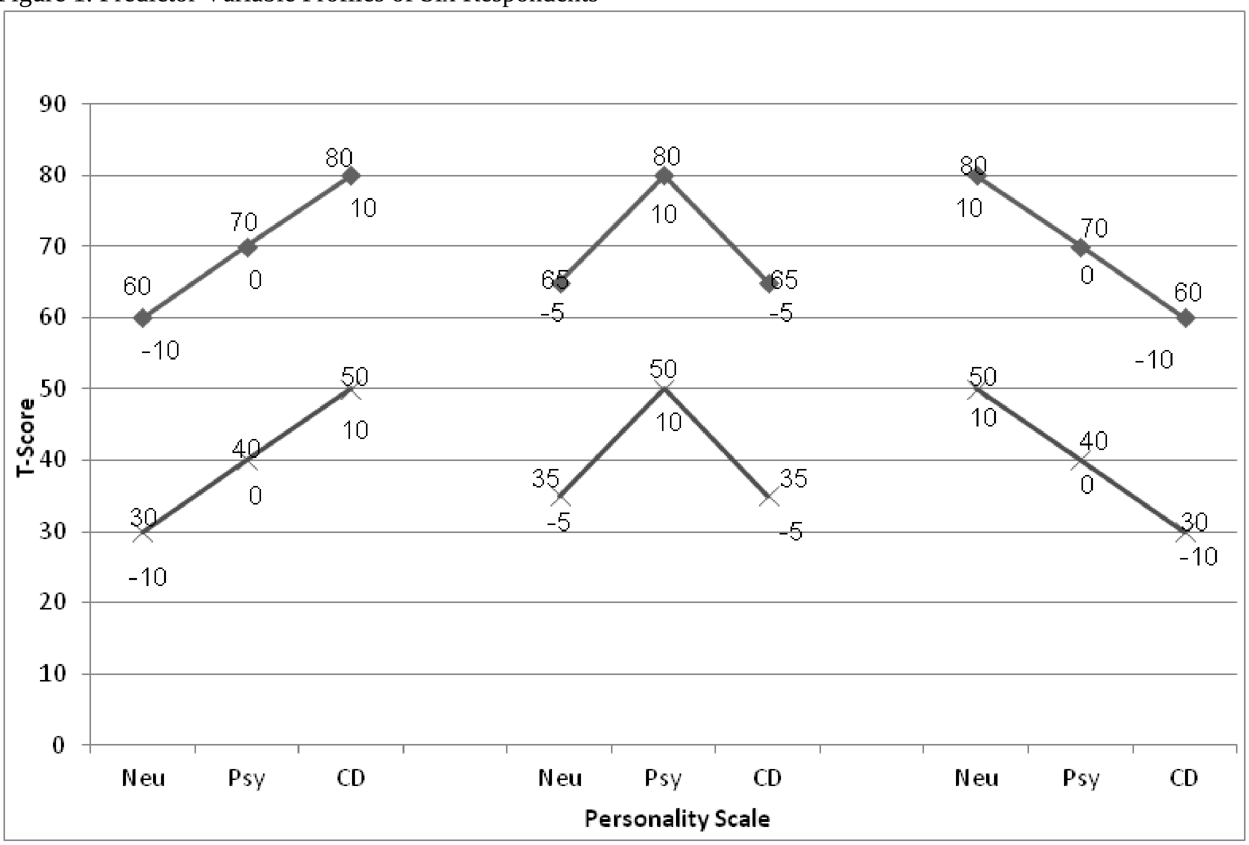
3.1 Definition

For clarity, we begin with a definition of a profile pattern. An individual profile pattern is the arrangement of scores in a respondent's vector of scores. Figure 1 shows six such profiles on a set of three personality predictor variables. Level of a profile refers to the height of that person's scores described by the mean of the scores in the profile:

– . Here, the pattern can be described by a vector of three contrast coefficients (one for each variable),

)} which represent the deviation of each score from the level . In Figure 1, the contrast coefficients describing each profile are shown below the profile. Each pair of profiles with the same visual shape has the same pattern vector of contrast coefficients. Two profiles are said to be mirror images if the contrast coefficients are of the same magnitude but opposite in sign. In Figure 1, the linearly increasing profiles on the left with pattern vectors (-10, 0, 10) are mirror images of the linearly decreasing profiles on the right with pattern vectors (10, 0, -10). If the three people in the upper portion of Figure 1 (with higher *Level* scores) have higher scores on some external criterion variable than do the three people in the lower portion, then we would say that individual differences in profile level are associated with variation in the criterion variable. On the other hand, if the two people in the middle of Figure 1 with the inverted V-shaped profiles have higher criterion scores than the other four, we would say that individual differences in profile patterns are associated with variation in the criterion variable. In addition to the profiles demonstrated, one can have individuals with fairly flat profiles, indicating very little scatter in their scores.

Figure 1. Predictor Variable Profiles of Six Respondents



3.2 Proof

1) General Regression Equation:

2) Prediction Equation:

3) (Reformulation for K predictors)

4) (Add / subtract constants)

5) (Combining terms)

Note that $b_k - \bar{b}$ provides the pattern of the optimal beta weights to the predictors given the criterion of interest. This pattern is optimal by least squares in predicting high values of the criterion.

6)] Expanding (a+b)(c+d)

7) Distribute summation sign Σ(A + B + C) = ΣA + ΣB + ΣC

8) Resolve deviation scores

9) (Adding a constant K times)

The original regression equation can be stated as a function of an implicit intercept, b_0, plus two new entities. Thus, the information in *K* predictors can be reduced to two. Moreover, these two new predictors have inherent meaning. The first is the numerator of a covariance between the predictors and the optimal regression weights. Hence, it measures the relationship between one's pattern of predictors versus that of the optimal beta weights - a match statistic. This match statistic can also be obtained from a prototypical profile obtained from some other process (e.g. expert judgment). The second entity in the re-expressed regression equation is a function of the mean of the predictors (elevation / level). Given the initial regression to obtain optimal regression weights relative to the criterion of interest allows our pattern (match statistic) to have criterion validity.

10) Since multiplying by a constant will affect only the regression weights, we can have a new regression equation that is equivalent to the first as:

11) Q.E.D

We now have a measure of *Level* plus an intra-observation covariance (match statistic) to the optimal pattern of the predictors as suggested by their beta weights. Here pattern is a combination of both shape and scatter.

4. Example

The example we use to illustrate potential uses of this technique explores the question of what is more important; the number of mathematics courses taken or the pattern of courses taken? Much of the previous literature confirms what we see in Table 1, that it is the types of math courses taken that matter. Table 1 gives data from the 2007 Digest of Education Statistics (Snyder, Dillow, and Hoffman, 2008) that shows cumulative math course-taking by ethnicity for high school graduates from 1982 to 2005. One can see a rise in courses taken over the years. In fact, the correlation between units taken and year is 0.979, indicating a consistent rise in courses taken over time. The course taking differences shown in Table 1, however, do not mirror typical performance differences we see. For instance, the White/Black difference in Carnegie units are small ranging from 0.13 favoring Whites (1994) to a 0.07 difference favoring Blacks (1990). The mean difference is less than 0.02 favoring Whites. Moreover, for three of the seven time points Black students actually took more courses (1990, 1998, and 2005). Given that this is not consistent with achievement differences that we see, we suspect that number of math courses may not be the metric we want to measure quality of mathematics course taking. While most of the previous literature agrees with this assertion, previous studies that claim to explore course taking patterns often fall short, settling for number of courses, number of advanced courses, highest course, etc. Given our technique we can truly discuss patterns of course taking.

Table 1: Average Carnegie Units Earned in Mathematics

Year	1982	1987	1990	1994	1998	2000	2005
Am Indian	2.35	2.98	3.04	3.11	3.10	3.29	3.53
Asian	3.15	3.71	3.52	3.66	3.62	3.96	3.90
Black	2.61	2.99	3.20	3.23	3.42	3.54	3.71
Hispanic	2.33	2.81	3.13	3.28	3.28	3.42	3.49
White	2.68	3.01	3.13	3.36	3.40	3.56	3.69
Total	2.63	3.01	3.15	3.33	3.40	3.56	3.67

From Snyder, T. D., Dillow, S. A. & Hoffman, C. M. (2008). Digest of Education Statistics, 2007 (NCES 2008-022). National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education. Washington, DC (Table 140). Retrieved from: Table 140 http://nces.ed.gov/programs/digest/d07/tables/dt07_140.asp.

For our example we use data from the National Educational Longitudinal Study of 1988 (Curtin, Ingels, Wu, and Heuer, 2002). Eighth-graders in 1988 were followed longitudinally. The study uses data from the base year (1988) when the students were at the beginning of their high school career and second follow-up (1992) when the students were high school seniors. In addition to providing a wealth of demographic information over a period of time for a base sample of 25,000 students in 1988, the NELS survey also gathered test data. Students were tested in four learning areas (reading, mathematics, science, and social studies) at three time points 1988, 1990, and 1992. The data also has high school transcript information for most of the students. The sample for the current study includes students for whom there were at least one transcript entry per year for four grades (9 - 12) and who also had valid math achievement data for both 1988 and 1992. Using these restrictions, the resulting sample size was 10,250.

All analyses employed the transcript weight, a weight designed to make the weighted sample of students for whom transcripts were collected representative of the national population of high school seniors in 1992. Since the data were collected from students sampled within schools violating the assumption of independence, regular standard errors are inappropriate (Kish, *1965).* Thus, all standard errors for statistical tests were computed assuming half as many subjects (design effect of 2). Standard practice suggests using hierarchical linear modeling (Raudenbush & Bryk, 2002) or re-sampling techniques such as "jack-knifing" to produce appropriate standard errors for such cluster samples. A discussion of "Design Effects and Approximate Standard Errors" can be found in NCES's user's manual (Curtin, et. al., 2002). As one samples more subjects from the same cluster, the design effect increases. The use of a design effect has precedence as seen in Hoffer (1997). His work in the area of student achievement likewise assumed a design effect of two (see notes to his Table 1).

The course content variable adopted here has advantages over previous approaches. First, it constitutes a description of the students' course content in terms of a continuous variable that accounts for all of the coursework, not just the highest level of coursework. Second, the content variable is suitable for analysis as a continuous variable in readily available software. Third, and most importantly, our course content variable, coupled with an amount of coursework variable, will fully capture the complete relationship between available coursework information and achievement and account for all variation in achievement. By using the amount and content variables described below, one does not risk mis-specifying the variables in a way that leads to underestimation of the variation in achievement that can be accounted for by the coursework information. We explicitly quantify course-taking as our match statistic. This is justifiable since it quantifies the match between courses taken and math achievement as measured by the tests.

The procedure begins by describing a taxonomy of math courses that fully captures the pattern of courses taken by individual students. Davenport, Davison, Kuang, Ding, Kim, and Kwak (1998) used multidimensional scaling to develop a reasonably concise, but comprehensive taxonomy for math courses. This taxonomy consists of several prototypical course sequences. A prototypical course sequence is an empirically derived set of courses that are taken by a significant subset of students. If students who take course "A" are more likely to take courses "B" and "D" as well, then courses "A", "B", and "D" will define a prototypical course sequence since a significant number of students take each of these three courses. Any other sets of courses with elevated probabilities of being taken by a substantial number of students will also emerge as a prototypical course sequence. The final taxonomy places *56* math courses in the Classification for Secondary School Courses (CSSC) into seven identifiable course sequences plus an "Other" category. The CSSC course titles used by many national surveys and assessments are described in Legum, Caldwell, Goksel, Haynes, Hynson, Rust, & Blecher (1993).

The course categories that emerged are as follows. Functional courses, at the lowest end of math literacy, represent survival skills in math. Basic courses are the minimal courses required for general math literacy. Preformal courses are terminal courses for some students, but provide background for other students who take more advanced courses. An Algebra sequence is composed of an Algebra 1 course given in two parts over two years. The Standard sequence consists of Algebra 1, Algebra 2, and Geometry and is the minimal set of math courses for a student on an academic track. Unified courses represent a different packaging of algebra and geometry concepts where topics are presented in an integrated manner. Courses in the Advanced sequence are usually taken by students preparing for college. Finally, the Other category contains courses not easily interpretable as part of any common high school course grouping or special offerings (oftentimes unique to a small number of schools and not taken frequently by students).

High school transcript data were used to obtain course information for each student. First we computed the number of Carnegie units earned for all math courses together as well as the number of Carnegie units earned in each of the eight course categories. The dependent variable, math achievement, was measured at base-year (1988 when the students were in 8th grade) as prior achievement and at the second follow-up (1992 when the students were in 12th grade). Use of prior achievement as a covariate is intended to statistically control for all other potential prior differences in the students; whether these differences are demographic, academic, etc. prior to high school.

Table 2 shows results of regressing senior math achievement onto the number of courses taken in the eight course categories. These categories accounted for *57.4%* of the variation in senior math achievement. Raw regression weights are in column 2 followed by the modified standard errors of the regression coefficients assuming a design effect of 2 in column 3. The t values (based on these modified standard errors) follow in column 4. All t values are significant with the exception of Algebra. The *Course Pattern Coefficients* are the un-standardized regression coefficients expressed as a deviation about the mean of the regression coefficients for the eight course categories. This latter index follows from the definition of pattern as what remains after elevation is accounted for. These coefficients specify the *Course Pattern* as a set of within person contrast coefficients, and like contrast coefficients in ANOVA, the coefficients sum to zero yielding both positive and negative values. Hereafter, they are called the *Advanced Pattern* coefficients because more advanced course categories have the higher coefficients. These weights map well with our expectation that students taking higher level math courses are more apt to score higher on math tests. Figure 2 has a plot of the advance course pattern. This figure clearly shows that we now have a pattern of course-taking that relates to mathematics achievement; with more advanced courses having more positive weights.

Table 2: Regression Coefficients for Predicting Senior Math Achievement from the Number of Carnegie Units in Eight Course Categories

	Regression Weight	Standard Error	t-value	Course Pattern Coefficient
Intercept	45.30	0.36	125.96	
Functional	-3.59	0.69	-5.20	-4.27
Basic	-3.02	0.30	-10.09	-3.70
Preformal	-2.32	0.16	-14.50	-3.00
Algebra	0.09	0.26	0.34	-0.59
Standard	2.38	0.12	19.94	1.70
Unified	2.67	0.19	13.91	2.00
Advanced	6.28	0.14	45.65	5.60
Other	2.94	0.22	13.56	2.26

R^2 = 57.4%. Standard errors are modified based on a design effect of 2.
All regression weights are significant at $p < 0.01$ with the exception of Algebra.

For each student, *Content* is computed as the covariance of the number of CUs taken in each course category versus the regression weight for that category (quantification of course content). A student receives a high *Content* score if their course taking reflected higher numbers of CUs for more advanced math courses and lower numbers of CUs for least advanced courses. Students with negative *Content* scores take most of their coursework in less advanced categories, ones with negative coefficients. Table 3 illustrates *Amount* and *Content* values for four students. Subjects 1 and 2 differ on amount of coursework (3 versus 6). While they take courses from the same categories, Standard and Advanced, Subject 2 takes more of these classes. Thus, Subject 2's *Content* score is higher, since (s)he had relatively higher CUs for the more advanced courses, matching the optimal course pattern slightly better. In contrast, Subject 3 took more low level courses and has a pattern that is a mirror image of the optimal and thus a negative *Content* score. Subject 4 had a flat pattern, taking one course each of the Basic, Standard, and Unified sequences (with no advanced course-work). Note, for all subjects the Math 12 score is in the same rank order as the *Content* score, again showing the consistency of our pattern statistic with the criterion. This is not true for *Amount*.

Figure 2: Advanced Mathematics Pattern Coefficients

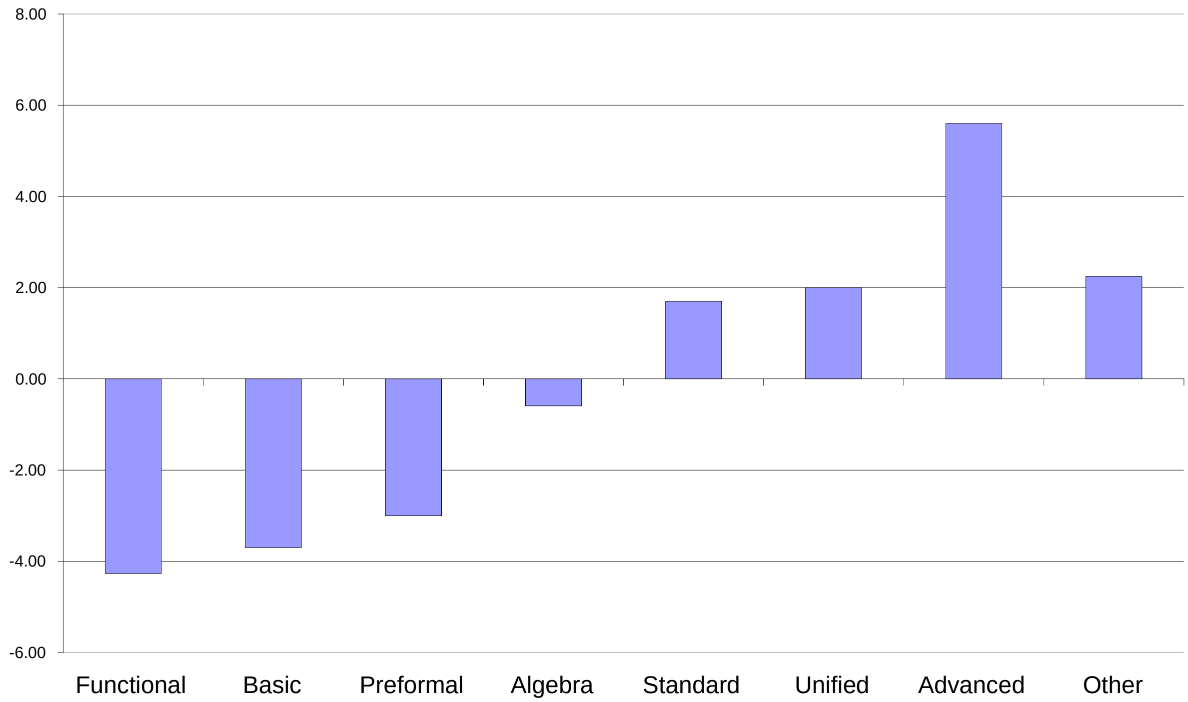


Table 3: Sample Results: Number of Carnegie Units in Each Category and Other Scores for Four Students

Category	Sub 1	Sub 2	Sub 3	Sub 4
Functional			2.50	
Basic			1.50	1.00
Algebra				
Standard	1.33	2.00		1.00
Unified				1.00
Advanced	1.67	4.00		
Amount	3.00	6.00	4.00	3.00
Content	1.44	3.23	-2.30	0.00
Math 8	71.62	69.04	38.38	41.99
Math 12	66.31	67.58	35.13	45.44

Correlations among prior math achievement (8th grade math test), senior math achievement (our primary dependent variable), total number of Carnegie units earned in math (Course *Amount*), and Course *Content* are shown in Table 4. Three findings from this analysis are useful in understanding later results. First, both prior achievement and senior achievement are correlated with coursework (*Amount* and *Content*), but senior achievement is more highly correlated with coursework than is prior achievement. Second, *Content* and *Amount* are correlated ($r = .56$). In general, students who took more courses show a pattern of more advanced courses. Third, *Content* was more highly correlated with both math tests than *Amount*. Finally, the single variable, *Content* ($r = .76$, $r^2 = 57\%$), accounted for as much variation in the 12^{th} grade math test as did virtually all eight coursework variables (See the R^2 at the bottom of Table 2. Thus, *Level* adds little to the predictability of math achievement above and beyond *Pattern*).

Table 4: Correlations among Math Achievement and Courses Taken

	Math 8	Math 12	Level	Content
Math 8	1.00			
Math 12	0.82	1.00		
Level	0.36	0.47	1.00	
Content	0.66	0.76	0.56	1.00

Amount accounted for 22% of the variation in senior math achievement. *Content*, by itself, accounted for *57%*, more than twice as much as *Amount*. With respect to increments in R^2 for predicting senior achievement, *Amount* added virtually nothing (0.4%) to the variance predicted from *Content* alone. In contrast, *Content* added an additional *35%* to the percent of variance predicted from *Amount* alone. These results raise the possibility that *Amount* of coursework is associated with senior achievement largely because students taking more coursework often (but not always) progress to more advanced courses. As stated above, *Amount* and *Content* together account for the same variation in senior math achievement as do all eight math course category variables in Table 2.

Above we show the importance of *Content* over *Level* for math course taking as it relates to math achievement. We can extend the usefulness of these statistics by exploring subgroup differences as it relates to *Level* or *Patterns* of course taking. Table 5 shows the effect size of math achievement using Whites as the referent group. It gives the distance in pooled standard deviation units between the mean of the group in question versus White students. Cohen (1988) gives guidance on interpreting effect size values. Values less than 0.2 are small, 0.5 represents a medium difference, and 0.8 a large difference. All of the Asian means exceeded those for Whites. For the other three ethnic groups the White means were larger. Note that all of the effects for *Amount* were small. The very small effect of 0.02 for the difference between number of courses taken for Whites and Blacks matches results from a host of studies given above as exemplified by the results shown in Table 1. With the exception of *Amount*, all of the other effects for American Indians were large, meaning that there is a large discrepancy between their means and that for Whites on the other three variables. The moderate to large effects for *Content* for American Indians, Blacks, and Hispanics better mirrors the difference represented in the 8^{th} and 12^{th} grade test scores. Thus, it again appears that explaining performance differences can be more readily done with *Content* differences than differences of *Amount*.

Moreover, we can extend this procedure to a moderated regression to ascertain whether amount and pattern have the same relationship to achievement for the different subgroups. While this remains a topic of discussion for another time, I can say that the relationship of *Amount* is similar for Whites and all of the other ethnic groups. The relationship of *Content*, however, is more complex. Both Blacks and Hispanics needed an additional *Content* variable as the effect of *Content* on achievement was different for them than for Whites.

Table 5: Effect Size Differences Using Whites as the Reference Group

Ethnic	Math 8	Math 12	Amount	Content
Am. Indian	-1.10	-1.17	-0.14	-0.93
Asian	0.09	0.20	0.13	0.23
Black	-0.75	-0.81	-0.02	-0.47
Hispanic	-0.63	-0.57	-0.25	-0.41

5. Discussion

This paper gives a two-step regression procedure that identifies an optimal pattern of scores as it relates to an external criterion and then provides quantification of a match for individual observations. As seen with the given example, this approach can be used to answer fundamental research questions; whether it is better to use amount or content of courses taken. Note that this approach can be used to parse *Level* and *Pattern* in a host of studies where one or the other is expected a priori to be more predictive. Note, too, that this approach can be used with a binary criterion. Culpepper (2009) generalized the analysis based on a multilevel, nonlinear model for dichotomous data. In application, the method has been used to study patterns in a wide variety of content areas including: high school coursework patterns as shown here, patterns of personality scores associated with vocational interests (Dilchert, 2007), patterns of IQ scores that distinguish types of disability (Chan, 2006), informal role patterns associated with team effectiveness (Coughlin, 2010), and personality score patterns associated with work behavior (Shen, 2011). There are also current efforts to adapt this approach to moderated regression to ascertain if there are differential effects of *Level* and *Pattern* via different subgroups and/or interactions of other potential predictors.

6. References

Chan, C-K. (2006). *An exploratory study of predictive validity and diagnostic utility of cognitive patterns on the Woodcock-Johnson Psychoeducational Battery---Revised (WJ-R) clinical database*. Minneapolis, MN: Unpublished Doctoral Dissertation, University of Minnesota.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Publisher.

Coughlin, D. E. (2010). *The Impact of Personality, Informal Roles, and Team Informal Role Configuration on Team Effectiveness*. West Lafayette IN: Unpublished doctoral Dissertation, Purdue University.

Cronbach, L. J., & Gleser, G. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*(6), 456-473.

Culpepper, S. A. (2009). A multilevel, nonlinear profile analysis model for dichotomous data. *Multivariate Behavioral Research, 44*, 646 – 667.

Curtin, T. R., Ingels, S. J., Wu, S. and Heuer, R. (2002). *National Education Longitudinal Study of 1988: Base-year to Fourth Follow-up Data File User's Manual* (NCES 2002:323). Washington, DC: US Department of Education, National Center for Education Statistics.

Davenport, E. C., Jr., Davison, M. L., Kuang, H., Ding, S., Kim, S, and Kwak, N. (1998). High school mathematics course-taking by gender and ethnicity. *American Educational Research Journal, 35(*3), 497-514.

Davison, M. L., & Davenport, Jr., E. C. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological Methods*, *7*(4), 468-484.

Dilchert, S. (2007). Peaks and valleys: predicting interest in leadership and managerial positions from personality profiles. *International Journal of Selection and Assessment*, *15*, 317 – 334.

Hoffer, T. B. (1997). High school graduation requirements: Effects on dropping out and student achievement. *Teachers College Record, 98*(4), 584-607.

Kish, L. (1965). *Survey sampling.* New York: John Wiley & Sons.

Legum, S., Caldwell, N., Goksel, H., Haynes, J., Hynson, C., Rust, K., & Blecher, N. (1993). *The 1990 high school transcript study data file user's manual.* Los Angeles: Westat, Inc.

Maxwell, S. E. and Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, N.J.: Lawrence Erlbaum.

Meehl, P. E. (1950). Configural scoring. *Journal of Consulting Psychology, 14*, 165-171.

Raudenbush, S. W. & Bryk, A. 5. (2002). *Hierarchical linear models* (2'' ed.). Thousand Oaks: Sage.

Shen, W. (2011). *The Application of a Person-oriented Criterion-related Configural Approach to the Relationship Between Personality Traits and Work Behaviors*. Minneapolis, MN: Unpublished doctoral dissertation, University of Minnesota.

Snyder, T. D., Dillow, S. A. & Hoffman, C. M. (2008). Digest of Education Statistics, 2007 (NCES 2008-022). National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education. Washington, DC (Table 140). Retrieved from: Table 140 http://nces.ed.gov/programs/digest/d07/tables/dt07_140.asp.